PB86-163854

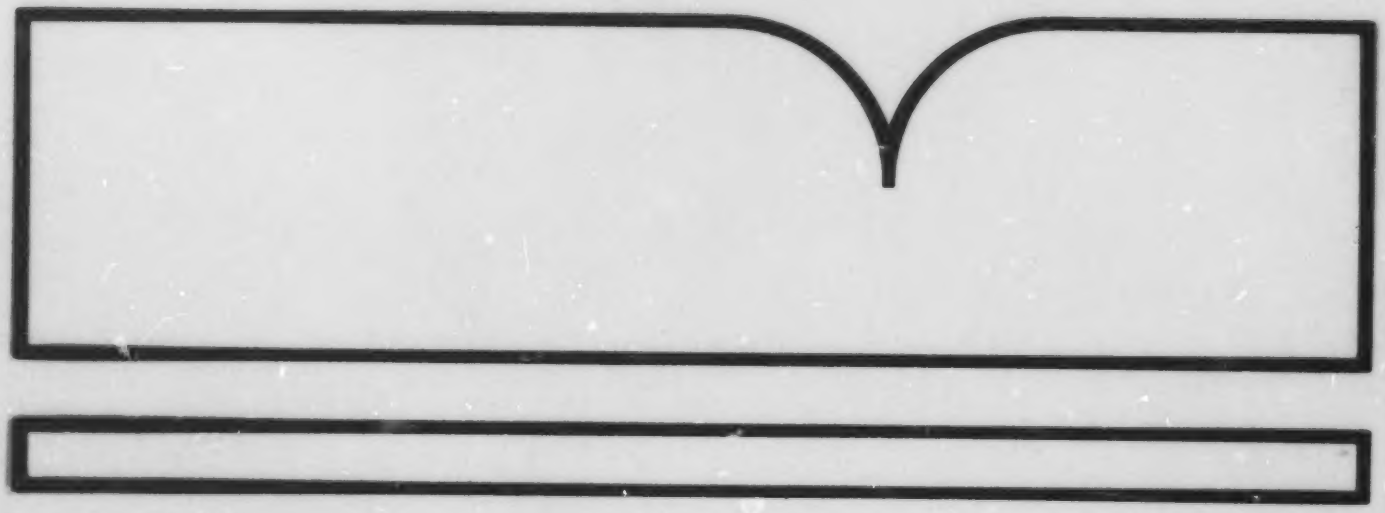
# Monte Carlo Studies of the Classifications Made by Nonparametric Linear Discriminant Functions. 2 Effects of Nonideal Data

Pennsylvania State Univ., University Park

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

1985



U.S. Department of Commerce
National Technical Information Service
NTIS

**TECHNICAL REPORT DATA**
*(Please read Instructions on the reverse before completing)*

| | | |
|---|---|---|
| 1. REPORT NO. EPA/600/J-85/319 | 2. | 3. PB86-163854 |
| 4. TITLE AND SUBTITLE: Monte Carlo Studies of the Classifications Made by Nonparametric Linear Discriminant Functions. 2. Effects of Nonideal Data. (Journal Version) | | 5. REPORT DATE: 1985<br>6. PERFORMING ORGANIZATION CODE |
| 7. AUTHOR(S): T. R. Stouch, P. C. Jurs | | 8. PERFORMING ORGANIZATION REPORT NO. |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: Pennsylvania State University, University Park, PA 16802 | | 10. PROGRAM ELEMENT NO. CCSL1A<br>11. CONTRACT/GRANT NO. CR811037 |
| 12. SPONSORING AGENCY NAME AND ADDRESS: Health Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711 | | 13. TYPE OF REPORT AND PERIOD COVERED: Journal Article<br>14. SPONSORING AGENCY CODE: EPA - 600/11 |

15. SUPPLEMENTARY NOTES: Ref: Journal of Chemical Information and Computer Sciences, 25(2):92-98, 1985

16. ABSTRACT

Recently, the levels of correct classifications due to chance that were attainable by nonparametric linear discriminant functions (NLDFs) were studied. That previous work dealt with easily generated, idealized data. Because of this, the application of those results to actual studies using nonideal data may not be warranted. The studies reported here analyze the effects of zero values, indicator values, and multicollinearities: variations that occur in actual data and that could affect the levels of random classifications. Three structure–activity relationship studies that were performed with NLDFs are also examined.

17. KEY WORDS AND DOCUMENT ANALYSIS

| a. DESCRIPTORS | b. IDENTIFIERS/OPEN ENDED TERMS | c. COSATI Field/Group |
|---|---|---|
| | | |

| 18. DISTRIBUTION STATEMENT | 19. SECURITY CLASS (This Report) | 21. NO. OF PAGES |
|---|---|---|
| RELEASE TO PUBLIC | UNCLASSIFIED | 8 |
| | 20. SECURITY CLASS (This page) UNCLASSIFIED | 22. PRICE |

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

EPA/600/J-85/319
JOURNAL ARTICLE

Reprinted from Journal of Chemical Information and Computer Sciences, 1985, *25*, 92
Copyright © 1985 by the American Chemical Society and reprinted by permission of the copyright owner.

NTIS is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

# Monte Carlo Studies of the Classifications Made by Nonparametric Linear Discriminant Functions. 2. Effects of Nonideal Data

TERRY R. STOUCH and PETER C. JURS*
The Pennsylvania State University, University Park, Pennsylvania 16802

Received December 20, 1984

Recently, the levels of correct classifications due to chance that were attainable by nonparametric linear discriminant functions (NLDFs) were studied. That previous work dealt with easily generated, idealized data. Because of this, the application of those results to actual studies using nonideal data may not be warranted. The studies reported here analyze the effects of zero values, indicator values, and multicollinearities: variations that occur in actual data and that could affect the levels of random classifications. Three structure–activity relationship studies that were performed with NLDFs are also examined.

Discriminant functions can be visualized as surfaces that divide a data space into different regions. The aim of this method of pattern recognition (PR) is to divide the data space into regions of significance. For example, in a structure–activity relationship (SAR) study the data space would be populated by points, often referred to as patterns, which represent compounds with interesting biological activity. A useful discriminant would divide the data space into regions

0095-2338/85/1625-0092$01.50/0 © 1985 American Chemical Society

that contained compounds of differing activity. The activity of new compounds could be predicted by the side of the surface on which a new compound's data point existed. Discriminants are useful means of reducing the complexity of some problems, because once a useful discriminant is found a single vector representing this surface could be used in place of the mass of experimental data that was used to generate it.

Linear discriminants are linear combinations of the variables used in the study and represent the simplest surfaces. Two-dimensional linear discriminants are straight lines separating two regions of two-dimensional space. A three-dimensional linear discriminant is a plane. Although no visualizations are possible in higher dimensions, the mathematical processes used to define the surfaces can be used with any number of dimensions. Parametric methods of discriminant generation require assumptions concerning the distribution of the data (usually multivariate normal) and generate the discriminant on the basis of the statistics of the estimated probability density function. Nonparametric methods use the raw data to generate the discriminant and require no assumptions concerning the distribution of the data.

The quality of a discriminant is often judged by the percentage of correct classifications that it provides. If a discriminant can correctly classify a large percentage of the data points correctly, it may indicate that the variables used in the study contain information pertinent to the properties that define the classes. In SAR studies this would mean that the variables contained information relating to the activity or lack of activity of the compounds in the study. Such a discriminant could presumably be used to predict the activity of new, untested compounds that were not used to generate the discriminant.

As with any mathematical method, there are some limitations to nonparametric linear discriminant functions. It has been known for some time that the probability of achieving 100% correct classification of a set of points, divided into two classes, increases as the number of variables used in the study increases.[1-3] For example, if the number of variables, $d$, equals the number of patterns, $N$, 100% correct classification is assured. This success is due only to the mathematics of the problem and is independent of any information contained in the variables. If the number of variables is large relative to the number of patterns, even a discriminant that provides completely correct classifications can be useless for prediction.

The variable to pattern ratio ($d/N$) is even more important in assessing discriminants that support less than 100% correct classification. Recently, we have studied such classifications in detail.[4] The relationship between $d/N$, the class size, and the level of correct but random classification was examined. We found that for a given $d/N$ there is some probability that a discriminant could be found that would afford 70%, 80%, 90%, 100%, etc. correct classification due solely to chance. These classifications were due only to mathematical artifacts and probabilistic considerations and could not be due to information contained in the data. The results from that study are summarized, in part, in Figure 1 (Figure 6 from reference 4). This shows the mean random correct classifications as a function of the $d/N$ ratio for two class problems of equally sized classes. For example, a study that consisted of 100 patterns, split 50/50 between two classes, and defined by 30 variables could, on the average, be 90% classified by a linear discriminant even if there were no class information contained within those variables. A 75% classification could be achieved if only 10 variables were used.

In that previous study, the data used were numbers generated by a random-number generator. The variables were continuous, of defined distribution, and contained a nonzero value for each observation. Also, they were largely uncorrelated. These data were quite ideal compared to the data that is commonly used in actual studies. Real data may contain discrete, "indicator" values instead of continuous variables. In real data, predicate variables are sometimes used, which lead to zero values in the variable. Such variables may indicate the presence or absence of key structural features. In a real study, the distribution of the variables might be unknown, or at least not consistant between the variables used in the study. Finally, real variables are seldom uncorrelated, and collinearity may be high. For example, molecular weight and volume are highly correlated structural properties.

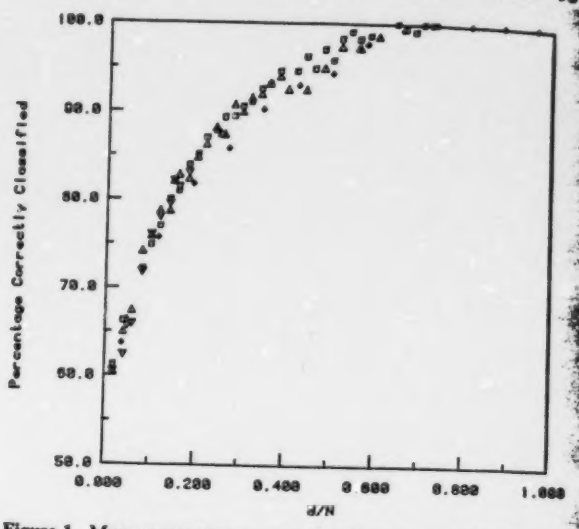



**Figure 1.** Mean percentage correct classification vs. $d/N$ (Figure 6 from reference 2; reprinted with permission. Copyright 1980 Springer-Verlag]: (diamonds) 26-pattern study; (squares) 50-pattern study; (triangles) 100-pattern study; (inverted triangles) 200-pattern study.

The data of the previous study were used because they were well characterized and easy to generate. Because of the differences between those data and "real" data, we did not know how the random classification levels determined in that previous work would apply to an actual study performed with data with the "deficiencies" listed above. We were especially interested in the effects of multicollinearity. When two variables are highly correlated, some of the variation in each variable is explained by the other. Highly collinear, high-dimensional data can often be explained by only a few noncollinear, orthogonal variables. This is the basis for principal component plots by which the largest part of the variance of a data matrix can be plotted in lower dimensional plots that allow convenient visual examination of the data. This is possible because multicollinearity has the effect of reducing the effective dimensionality of a data space. Since the level of random correct classifications for a problem is dependent, in part, on the number of dimensions used, multicollinearity might reduce the levels of these fortuitous correct classifications.

The aim of the work reported in this paper is to determine if these data deficiencies cause any substantial changes in the levels of random correct classifications reported in the previous work. Such changes would prevent the application of those results to actual studies.

## METHODOLOGY

Except for the alterations in the variables in order to simulate the various deficiencies, these studies were performed in a manner identical with that of the previous studies.[4] The variables used in these studies were vectors of random numbers generated by the algorithm of Schrage.[5] Gaussian distributions were approximated by applying the Central Limit Theorem.[6]

These were appropriately altered in order to show the desired deficiencies. In studies of predicate variables, the appropriate levels of zero values for each class were obtained by setting randomly chosen observations to zero. Indicator values were obtained by suitable multiplication and truncation of the continuous, 0.0–1.0, random variables to provide values from "1" to the number of indicator values.

Data matrices containing multicollinearities were generated by subtracting multiples of random vectors from a core random vector to form new, partly collinear, vectors. These new vectors and the core vector formed the data matrix. The size of the multiplicative constant determined the degree of collinearity between the core and new vectors. A large constant caused a large random factor to be removed from the core vector and resulted in a low correlation. A small constant resulted in a high correlation. The size of the desired matrix determined the number of new vectors that were calculated. The multiplicative constants were empirically determined as those which yielded the desired levels of collinearities and multicollinearities as reported by the collinearity indicators described under Results and Discussion.

Once sets of the appropriately "degraded" variables were obtained, they were submitted to the linear discriminant generating schemes used previously. The data matrices were analyzed with a series of algorithms including a perceptron[3] and an adaptive least-squares method.[7] This scheme was chosen for its speed of execution and its efficiency at developing a discriminant that could provide high classifications. The programs were run in an identical manner for all of the individual runs. The only difference between the runs was that the data matrices were formed from different variables.

Several of the routines were iterative, and some required the specification of adjustable error-correction parameters. The iteration limits and values of the parameters were determined from past experience with the algorithms. They were chosen in order to partially automate the studies and reduce the expenditure of computer time. If the iteration limit was increased or the adjustable parameters varied differently within a discriminant-generating run, the classification results may have improved. Past experience indicated that the values that were used in this study would usually achieve high correct classifications in a short time. There is no guarantee, however, that the results might not have improved given more freedom. In all cases a two-class problem was used, simplifying the problem as was done previously.

There are an infinite number of possible combinations of number of patterns, number of variables, collinearity, levels of zero values, and indicator values. Obviously, an all-encompassing study could not be performed. Instead, we first examined the effects of the isolated deficiencies, and we then examined the effects of some combinations by simulating three previously reported SAR studies.

These studies were formulated to determine if there were substantial changes in the classifications provided by "real" data as opposed to those of the idealized data that were studied previously. They were not performed in enough detail to determine small differences. More precision would have required considerably more resources and would have been of little value for two reasons. First, Figure 1 shows the mean of the random correct classifications at each $d/N$ ratio. The actual classifications had fairly large ranges. For the studies with a small number of patterns, the range was up to 10%–12%. Second, due to the iterative nature of the PR routines, there was no guarantee that any given classification results were the highest that a given data matrix could support. Any actual study will have only one "best" result. The important feature of these studies is the general range of the classifications—the levels that could be achieved. More precise identification of changes in the random classifications would have been mathematically interesting but of little practical value.

**Table I. Effects of Multicollinearities on Studies with 100 Observations**

| no. of variables | multicollinearity indicators | | | % correct classifications | | |
|---|---|---|---|---|---|---|
| | first PC[a] | 99%[b] | CN[c] | mean | SD | no. of runs |
| 10 | 9 | 11 | 1 | 75.1 | 2.6 | 20 |
| | 78–81 | 10 | 16–24 | 75.0 | 2.3 | 20 |
| | 80–85 | 6–7 | 57–80 | 74.7 | 2.4 | 19 |
| | 91–94 | 4–5 | 112–160 | 74.6 | 2.7 | 20 |
| 20 | 5 | 21 | 1 | 85.3 | 3.1 | 20 |
| | 77–79 | 18–19 | 30–40 | 84.8 | 3.0 | 20 |
| | 79–83 | 13–14 | 100–130 | 83.6 | 3.8 | 20 |
| 30 | 3 | 31 | 1 | 90.1 | 2.0 | 18 |
| | 76–78 | 26–27 | 50–70 | 90.3 | 1.7 | 9 |
| | 75–79 | 20–21 | 150–200 | 89.2 | 2.1 | 18 |

[a] Percentage of total variance accounted for by first principal component. [b] Number of principal components needed to account for 99% of the variance. [c] Overall condition number.

All the studies were performed on the Chemistry Department PRIME 750 computer. The software used to perform the studies was the ADAPT chemical software system, which has been described previously.[8]

## RESULTS AND DISCUSSION

In the first series, we examined the effects of the isolated deficiences. We could not examine all the $d/N$ ratios so we examined several $d/N$ ratios for problems of two different sizes, one with 100 patterns and one with 26 patterns. In both cases, the patterns were equally distributed between two classses. We chose these two sizes because they were the same as two of the studies done in the previous random classification work, because we felt them to be representative of the sizes of several studies already in the literature, and because they were of a size to be computationally convenient. We felt justified in using only these two sizes because of the close agreement between studies involving different class sizes as shown in Figure 1.

**Collinearity.** Our primary interest was in the effect that collinearity might have in the levels of random classifications. Collinearities in a data matrix are not simple to measure or report.[9,10] They may exist between two variables, between one variable and several others, or as complex relationships between many of the variables. The collinear matrices that were used in the studies reported here contained a combination of these. Three measures were used to indicate the degree of the multicollinearity: (1) the percentage of the total variance in the matrix accounted for by the first principal component, (2) the number of principal components that were required to account for 99% of the variance, and (3) the overall condition number of the matrix. The overall condition number is the square root of the ratio of the largest eigenvalue to the smallest eigenvalue. A large condition number indicates that a large percentage of the variance is present in the first principal component and that a small amount is present in the last. This quantity has been suggested as an indicator of collinearity by several authors.[9,10] Belsely, Kuh, and Welch have suggested that a value above 30 suggests the presence of high levels of collinearity and that a value above 100 indicates severe collinearity.[9] Since the discriminant-generating routines used in this work all add an extra dimension to the data matrix, this dimension was also included in the collinearity screening. Several levels of collinearity were generated for each $d/N$ ratio investigated. For each level, 9–20 individual matrices were generated.

Table I shows the results for the study using 100 patterns. Sets of 10, 20, and 30 variables were used as the data matrix

**Table II.** Effects of Multicollinearities on Studies with 26 Observations

| no. of variables | multicollinearity indicators | | | % correct classifications | | |
|---|---|---|---|---|---|---|
| | first PC[a] | 99%[b] | CN[c] | mean | SD | no. of runs |
| 5 | 17 | 6 | 1 | 86.7 | 4.1 | 15 |
| | 80–86 | 5 | 9–13 | 86.7 | 4.1 | 15 |
| | 87–94 | 3 | 50–101 | 84.3 | 6.5 | 15 |
| 10 | 9 | 11 | 1 | 95.4 | 3.5 | 10 |
| | 77–82 | 9 | 20–34 | 95.8 | 2.8 | 10 |
| | 89–95 | 5–6 | 112–160 | 93.9 | 5.5 | 10 |
| 15 | 6 | 16 | 1 | 98.7 | 3.0 | 12 |
| | 76–80 | 12–13 | 34–71 | 99.0 | 2.4 | 12 |
| | 89–94 | 7–8 | 127–208 | 97.7 | 3.2 | 10 |

[a] Percentage of total variance accounted for by first principal component. [b] Number of principal components needed to account for 99% of the variance. [c] Overall condition number.

that was submitted to the PR routines. For each dimensionality, the first line shows the classifications for completely orthogonal data. Each succeeding line shows increasing levels of collinearity as noted by the collinearity indicators. From the classification levels shown, collinearity in the data matrix appears to have little effect on the random classification levels. The most severely collinear data matrix had condition numbers much greater than 100 and had 99% of the variance explained in a number of principal components, which was far fewer than the number of variables in the data matrix. The successful classifications for these data matrices were not very different from those for the completely orthogonal data, which had an overall condition number of 1 and which required all of the dimensionality to explain 99% of the variance.

Table II shows similar results for the studies using 26 patterns. Even with fairly severe collinearity, the classifications closely paralleled those that were reported in the previous work. In the very collinear data sets, the apparent dimensionality of the data was far less than the number of variables used in the study. The random classifications, however, appear to be unaffected by those levels of collinearity.

Extremely collinear data would be expected to affect classification results. As an extreme example, two completely collinear variables contain no more information than either of the variables alone. In that case, classifications should be identical regardless of whether one or both of the variables were used. None of the data matrices used in these studies contained such singularities. Such data would never be used in an actual study and would not be useable for those PR routines that require matrix inversion. The matrices used here were more typical of actual data that has been used in actual PR-SAR studies. This example however, should serve as a hypothetical extreme limit to the problems that collinearities could cause. Apparently, the collinearity that was generated here did not approach that extreme. However, in the most collinear cases it was more extreme than that which is typically encountered in actual studies. At reasonable levels collinearity appears to have little effect on the levels of random correct classifications and so would be expected to have little effect on the application of the results in Figure 1 to actual studies.

**Indicator Variables.** The effect of using indicator variables was investigated. Indicator variables often occur in multivariate analysis. These serve to place an observation within one of two or more groups. The data that we used previously were continuous and could take on an infinite number of values. Reducing this to a few, possibly only two, discrete values might have the effect of restricting the number of positions in the data space that an individual point could occupy.

As above, we generated series of variables with 26 and 100 observations, which were equally distributed among two classes. Each variable was filled with integer values from "1" to the number of indicator values desired. For two indicator values, all the observations had values of either 1 or 2. For four indicator values, all the observations had the values of 1, 2, 3, or 4. For each $d/N$ ratio investigated, several levels of indicator values were generated. As the number of values increased, the data approached the continuous variables used in the previous study.

**Table III.** Effects of Indicator Values on Studies with 100 Observations

| no. of variables | no. of indicator values | % correct classifications | | |
|---|---|---|---|---|
| | | mean | SD | no. of runs |
| 1 | continuous | 58.0 | 3.4 | 15 |
| | 8 | 56.2 | 1.8 | 5 |
| | 6 | 58.0 | 1.2 | 5 |
| | 4 | 58.4 | 2.1 | 5 |
| | 2 | 53.4 | 2.0 | 5 |
| 15 | continuous | 82.2 | 4.0 | 5 |
| | 8 | 80.8 | 3.4 | 5 |
| | 6 | 81.6 | 2.3 | 5 |
| | 4 | 81.0 | 2.7 | 5 |
| | 2 | 80.2 | 2.2 | 5 |
| 20 | continuous | 84.9 | 2.2 | 10 |
| | 8 | 84.6 | 1.7 | 5 |
| | 6 | 86.4 | 2.8 | 5 |
| | 4 | 85.7 | 1.8 | 5 |
| | 2 | 85.6 | 1.5 | 5 |
| 30 | continuous | 90.0 | 1.9 | 20 |
| | 8 | 89.2 | 1.1 | 5 |
| | 6 | 90.4 | 2.6 | 5 |
| | 4 | 90.0 | 1.6 | 5 |
| | 2 | 90.6 | 1.5 | 5 |
| 40 | continuous | 92.6 | 3.2 | 9 |
| | 8 | 92.4 | 3.6 | 5 |
| | 6 | 93.6 | 4.2 | 5 |
| | 4 | 92.2 | 1.1 | 5 |
| | 2 | 92.0 | 2.6 | 5 |

**Table IV.** Effects of Indicator Values on Studies with 26 Observations

| no. of variables | no. of indicator values | % correct classifications | | |
|---|---|---|---|---|
| | | mean | SD | no. of runs |
| 5 | continuous | 81.8 | 6.7 | 17 |
| | 8 | 76.2 | 5.0 | 5 |
| | 6 | 86.2 | 5.9 | 5 |
| | 4 | 83.1 | 3.4 | 5 |
| | 2 | 85.4 | 5.0 | 5 |
| 10 | continuous | 94.0 | 5.2 | 4 |
| | 8 | 93.8 | 5.2 | 5 |
| | 6 | 95.4 | 4.2 | 5 |
| | 4 | 96.2 | 2.7 | 5 |
| | 2 | 92.3 | 5.5 | 5 |
| 15 | continuous | 97.8 | 4.2 | 12 |
| | 8 | 100 | 0.0 | 5 |
| | 6 | 100 | 0.0 | 5 |
| | 4 | 100 | 0.0 | 5 |
| | 2 | 97.8 | 3.4 | 5 |
| 20 | continuous | 100 | 0.0 | 12 |
| | 8 | 100 | 0.0 | 5 |
| | 6 | 100 | 0.0 | 5 |
| | 4 | 100 | 0.0 | 5 |
| | 2 | 100 | 0.0 | 5 |

Table III shows the results for 100 patterns. For each different $d/N$ ratio, a decrease in the number of values within each variable did not decrease the classifications noticeably. Even when each variable contained only two possible values, no great changes in the random classifications were seen. The results for the 26-pattern studies (Table IV) also showed little or no deviation from the results obtained with "nondegraded" variables.

**Zero Values.** The final study of the isolated deficiencies was that of predicate variables. Several SAR studies, including those that are examined subsequently, used variables wherein

Table V. Effects Zero Values on Studies with 100 Observations

| no. of variables | % omitted values | % correct classifications | | |
|---|---|---|---|---|
| | | mean | SD | no. of runs |
| 1 | 0 | 58.0 | 3.4 | 15 |
| | 25 | 57.2 | 2.3 | 10 |
| | 50 | 55.7 | 2.1 | 10 |
| | 75 | 54.2 | 2.3 | 10 |
| 10 | 0 | 75.9 | 2.0 | 10 |
| | 25 | 72.9 | 3.1 | 20 |
| | 50 | 72.3 | 2.0 | 22 |
| | 75 | 67.6 | 2.6 | 20 |
| 15 | 0 | 82.2 | 4.0 | 5 |
| | 25 | 78.8 | 1.5 | 10 |
| | 50 | 77.0 | 1.6 | 10 |
| | 75 | 75.4 | 1.4 | 10 |
| 20 | 0 | 84.9 | 2.2 | 10 |
| | 25 | 82.8 | 2.7 | 10 |
| | 50 | 80.6 | 1.7 | 10 |
| | 75 | 79.5 | 2.8 | 10 |
| 30 | 0 | 90.0 | 1.9 | 10 |
| | 25 | 87.2 | 1.7 | 10 |
| | 50 | 86.8 | 2.2 | 10 |
| | 75 | 85.0 | 2.7 | 10 |
| 40 | 0 | 92.6 | 3.2 | 9 |
| | 25 | 89.7 | 2.7 | 10 |
| | 50 | 88.0 | 1.9 | 10 |
| | 75 | 88.3 | 2.1 | 7 |

Table VI. Effects Zero Values on Studies with 26 Observations

| no. of variables | % omitted values | % correct classification | | |
|---|---|---|---|---|
| | | mean | SD | no. of runs |
| 1 | 0 | 63.3 | 3.7 | 15 |
| | 25 | 62.7 | 4.5 | 13 |
| | 50 | 61.5 | 3.8 | 13 |
| | 75 | 58.9 | 1.8 | 13 |
| 3 | 0 | 75.7 | 7.4 | 16 |
| | 25 | 72.1 | 3.4 | 8 |
| | 50 | 69.7 | 4.8 | 8 |
| | 75 | 65.5 | 4.1 | 8 |
| 5 | 0 | 81.2 | 6.7 | 17 |
| | 25 | 80.4 | 5.8 | 13 |
| | 50 | 79.0 | 4.9 | 13 |
| | 75 | 71.3 | 4.3 | 13 |
| 10 | 0 | 94.0 | 5.2 | 4 |
| | 25 | 90.0 | 4.0 | 13 |
| | 50 | 88.8 | 4.3 | 13 |
| | 75 | 83.7 | 3.6 | 13 |
| 15 | 0 | 97.8 | 4.1 | 12 |
| | 25 | 97.6 | 4.0 | 13 |
| | 50 | 94.4 | 5.8 | 13 |
| | 75 | 91.7 | 4.4 | 13 |
| 20 | 0 | 100 | 0.0 | 12 |
| | 25 | 100 | 0.0 | 5 |
| | 50 | 97.7 | 5.1 | 5 |
| | 75 | 97.7 | 5.2 | 5 |

the presence of a nonzero value was predicated by the occurrence of some structural feature, often a functional group. If that group was not present in a structure, the value of the variable for that structure was zero. This value actually conveys information concerning that structure and so this situation deviates somewhat from situations where a measurement is missing for some less substantial reason. This might be considered to be a variant of the indicator value problem examined above. Numerically, this could have the effect of creating "holes" in the data matrix. These studies were performed to investigate the effect of such holes on the levels of random correct classifications.

The studies were run as before, for 26 and 100 patterns, for several different matrix sizes for each, for several levels of zero values for each of these, and for multiple runs at each level.

Table V shows the results for 100 patterns and Table VI those for 26 patterns. They show slight decreases in the levels of classifications as the levels of zero values increases. These decreases approach 10% when the level of zero values is 75%. The case of 75% zero values is an extreme case, however. All of the variables had only 25% of the observations coded by anything but zeros. At such an extreme there is a high probability that some of the randomly coded patterns will contain no nonzero values, a case that would never arise in an actual study. It is unlikely that data with 75% zero values would ever by used in an actual study. Even if one variable in a study was so extreme, it is unlikely that all the variables would be of such poor quality. For levels of zero values below 50%, however, the decrease in random correct classifications, while significant, is not great and the ranges of the classifications of the idealized and the degraded data overlap to a great extent.

**Combination of Effects.** From these results it is apparent that, except in extreme cases, the effects of the degradations are slight or nonexistant. In particular, even fairly extreme collinearity does not reduce the levels of random correct classifications. All of these effects were analyzed in the absence of the others, however. "Real" data often contains a combination of these effects, and it is conceivable that there may be synergistic effects between them. The next step in these studies was to generate sets with such combinations. Rather than trying to determine a few likely combinations of these effects—there are an infinite number of possibilities—we chose to simulate some actual SAR studies that have utilized nonparametric linear discriminants. In this way, we could examine combinations that not only could occur but have, in fact, occurred.

The first study that we simulated was that of the SAR of 213 antitumor 9-anilinoacridines (9-AA) reported previously.[11] This study involved the classification of these compounds as either having or not having antitumor potential. The discriminant generating stage of the study consisted of 153 active and 60 inactive compounds. The authors reported that 94% correct classification was possible with a set of 18 variables.

In the first stage of the simulations, sets of 18 continuous, uniformly distributed and completely coded variables were generated. A total of 153 patterns was placed in one class, and 60 were placed in a second, in order to simulate the class sizes used in the previous study. When these sets of variables were run through the same discriminant generating routine that was used in all the previous studies, classifications ranged from 79.8% to 82.2% (mean 81.5%, SD 1.0%, eight runs). These are the results that would be expected from appropriate extrapolation of the results in reference 4.

The variables used for these simulations differed substantially from those used in the 9-AA study, however. The variables used in the actual study contained many omitted values. There were two-leveled and multileveled indicator variables that represented the presence of various substructures. Some appeared to have a normal distribution, while others were highly skewed or appeared more uniformly distributed. A summary of this information is presented in Table VII.

New sets of variables were generated with the same number of observations and class assignments as above but with duplications of the predicate variables, indicators, distributions, and collinearities. In some cases the duplications were crude, but the results were interesting. A total of 10 sets of these variables yielded 88.7%–91.5% correct classifications (mean 90.4%, SD 0.87%, $n = 10$). This is an increase of almost 10% over the unaltered variables.

Since the previous studies indicated that the various degradations had little or no effect on classifications, further studies were performed in order to isolate the cause of the increased correct classifications. Another series of sets of variables was generated that duplicated only the indicator values present in the actual variables. Classifications for 10

**Table VII.** Characteristics of 9-Anilinoacridine Data

| index[a] | no. of indicator values | % nonzero values | | |
|---|---|---|---|---|
| | | total | class 1 | class 2 |
| 1 | 2 | 69 | 80 | 40 |
| 2 | 2 | 100 | 100 | 100 |
| 3 | normal[b] | 100 | 100 | 100 |
| 4 | normal | 100 | 100 | 100 |
| 5 | 2 | 33 | 32 | 37 |
| 6 | 5 | 21 | 24 | 13 |
| 7 | 5 | 23 | 25 | 18 |
| 8 | 2 | 16 | 19 | 10 |
| 9 | 5 | 27 | 29 | 20 |
| 10 | 4 | 19 | 22 | 12 |
| 11 | 3 | 42 | 46 | 32 |
| 12 | 4 | 63 | 74 | 35 |
| 13 | 3 | 32 | 35 | 25 |
| 14 | 6 | 69 | 60 | 92 |
| 15 | 2 | 26 | 22 | 35 |
| 16 | 5 | 100 | 100 | 100 |
| 17 | 8 | 100 | 100 | 100 |
| 18 | 3 | 100 | 100 | 100 |

[a] In the same sequence as the original study. [b] Normally distributed.

runs of these variables ranged from 81.2% to 83.1% (mean 82.7%, SD 0.7%). These results varied little from those of the unaltered variables.

More interesting results were obtained when variable sets were generated with no indicator information and only the information concerning the predicate variables. Classifications for eight sets were between 85.5% and 89.7% (mean 87.2%, SD 1.3%). The predicate variables appear to have caused the bulk of the increase in correct classifications. The zero values for these simulations differed from the studies in the first part of this section. The studies of the isolated effects of predicate variables contained the same levels of zeros in each of the two classes. In most cases, the percentages of zero values in the actual study were not constant between the two classes. In three cases, the difference in these values was greater than 30%, and six others had differences of greater than 10%. Furthermore, in all but two of the skewed variables the smallest class had the higher level of zero values. When the levels of zeros were adjusted to be the same in both classes, classifications dropped to around 80%. This held both when the percentage of omitted values was that of the greatest and that of the least coded class. Through further simulations, the bulk of the increase was found to be due primarily to those variables with the greatest differences in zero values between the two classes. In this case, the various "degradations" in the data not only did not decrease the levels of random correct classifications but served to increase them substantially.

It may be argued that these variables do contain some information. They may provide information on the presence or absence of certain properties, and therefore, criticism of the use of these variables in the 9-AA study may not be justified. Further evidence supporting the usefulness of the discriminants developed in the actual study was the reported prediction results. The authors reported that the discriminant was capable of correctly predicting the activity of up to 82% of some new compounds that were not used in the development of the discriminant. This indicates that there was information in the variables that pertained to biological activity. The 94% classifications could well have been real and based on the uncovering of an actual SAR and not due to chance. Note that the possibility of high random correct classifications does not preclude the possibility of actual clustering of the patterns. The classifications in the 9-AA study were higher than those that were obtained from the simulations. While it was a difference of only 4%, it may provide evidence of the uncovering of a true SAR. The results reported here are interesting in that they show that similar results could have been obtained by using appropriate random variables that would contain no information pertaining to biological activity.

**Table VIII.** Characteristics of *N*-Nitrosamine Data

| index[a] | no. of indicator values | % nonzero values | | |
|---|---|---|---|---|
| | | total | class 1 | class 2 |
| 1 | 4 | 100 | 100 | 100 |
| 2 | normal[b] | 100 | 100 | 100 |
| 3 | 13 | 100 | 100 | 100 |
| 4 | normal | 100 | 100 | 100 |
| 5 | 11 | 100 | 100 | 100 |
| 6 | 4 | 100 | 100 | 100 |
| 7 | 2 | 16 | 12 | 28 |
| 8 | normal | 100 | 100 | 100 |
| 9 | normal | 100 | 100 | 100 |
| 10 | 4 | 80 | 86 | 63 |
| 11 | 9 | 91 | 91 | 92 |
| 12 | 3 | 16 | 12 | 28 |
| 13 | 2 | 16 | 12 | 28 |
| 14 | 4 | 80 | 86 | 63 |
| 15 | normal | 80 | 86 | 63 |
| 16 | 2 | 8 | 6 | 13 |
| 17 | 3 | 48 | 51 | 36 |
| 18 | 10 | 48 | 51 | 36 |
| 19 | 3 | 17 | 14 | 26 |
| 20 | 2 | 7 | 5 | 13 |
| 21 | 4 | 17 | 14 | 26 |
| 22 | 3 | 17 | 14 | 26 |

[a] In the same sequence as the original study. [b] Normally distributed.

**Table IX.** Characteristics of *E. coli* Data

| index[a] | no. of indicator values | % nonzero values | | |
|---|---|---|---|---|
| | | total | class 1 | class 2 |
| 1 | 3 | 36 | 26 | 46 |
| 2 | 9 | 100 | 100 | 100 |
| 3 | normal[b] | 100 | 100 | 100 |
| 4 | 6 | 77 | 82 | 73 |
| 5 | normal | 38 | 22 | 53 |
| 6 | 3 | 12 | 10 | 15 |
| 7 | normal | 65 | 68 | 64 |
| 8 | 7 | 32 | 44 | 22 |
| 9 | normal | 41 | 38 | 44 |
| 10 | normal | 38 | 40 | 37 |

[a] In the same sequence as the original study. [b] Normally distributed.

The second simulated study was that of 112 carcinogenic and 38 noncarcinogenic *N*-nitrosamines reported by Rose and Jurs.[12] The authors reported 97% correct classification of these training set compounds with a set of 22 variables. Information pertinent to the simulations is reported in Table VIII. The classifications achieved for 22 unaltered random variables of 150 patterns assigned 38/112 to two different classes ranged between 87% and 92% (mean 88.71%, SD 1.80%) When the information in Table VIII was used to simulate the variables actually used in the study, classifications ranged between 90% and 94% (mean 91.8%, SD 1.40%). This is somewhat higher than that for the nondegraded data but not nearly as extreme as that for the 9-AA simulations. The difference in the levels of omissions between the two classes were not as extreme in this study; the greatest difference was 23%. Also, the high levels of zero values did not occur predominantly in one class. The variable set for this study contained a variety of types of variables with a variety of predicate variables and indicator values. Once again, there appear to be no joint effects that serve to dramatically lower the levels of random correct classifications.

The final actual study examined was that of 55 mutagenic and 50 nonmutagenic compounds as identified in an in vitro mutagen screening test employing the *Escherichia coli* WP2uvrA mutant. This SAR study is presented, in part, elsewhere.[13] The results of this study were a set of 10 structural variables and a discriminant that could classify 90%

of the 105 structures. The pertinent information for the simulations is presented in Table IX. Without the data deficiencies, 10 variables of 105 patterns split 50/55 between two classes yield classifications of between 71% and 80% (mean 75.5%, SD 3.41%). When the actual variable set was simulated, classifications ranged from 75% to 82% (mean 78.5%, SD 2.12%). This result is only slightly higher than that of the nondegraded data. Once again, the levels of omissions were not as great or as skewed as those in the 9-AA study. While these results appeared promising, prediction studies using 40 compounds that were not used in discriminant generation or in variable selection and development were essentially random. In this case, and for these prediction compounds, the variable set and discriminant were useless for prediction even though the classifications made by the discriminant were greater than those attributable to chance alone. The additional 10% of the classifications in the discriminant generating stage were found to be due to clustering of structurally similar compounds in the structurally very heterogeneous data set.

The significant point of these simulations is that combinations of the degradations did not decrease the random correct classifications. While these three cases do not begin to cover the infinity of possible combinations of multicollinearities, omissions, and indicators, they serve a more important function of duplicating some combinations that have actually occurred, unlike some of the extreme cases of omissions present in the first part of this section.

## CONCLUSIONS

The results presented here indicate that the levels of random correct classifications observed previously[2] for the idealized data can be applied with some confidence to studies where the data are not so ideal. Except in extreme cases, classifications for problems with the same numbers of patterns and variables varied little between the idealized data and data containing predicate variables, indicator values, and multicollinearity. In fact, the simulations of the 9-anilinoacridine study show that "skewing" of the data between classes can increase classifications. While this effect is easily seen here, this study was a simple case. Such skewing between classes may not always be so obvious. The effect could occur whenever one class was predominantly coded with one value. Such variables could potentially contain useful information; however, as seen here, they could also contribute to random classifications if improperly applied.

These results do not mean that nonparametric linear discriminants are not useful. They can play a useful role in data reduction. Also, when the ratio of variables to patterns is low, they can be used with some confidence as a means for determining the separability of the classes. A ratio of 1/20 or less would allow less than 70% random correct classifications. Results substantially higher than this could be treated with confidence. Discriminants should not be used as reliable indicators of the clustering of patterns in the data space when this ratio is high, however. A ratio of 1/10 would allow between 70% and 80% random correct classifications; a 1/3 ratio, between 85% and 95% random correct classifications. This does not mean that a relationship does not exist among the compounds at high dimensionality. As was noted previously, the fact that high levels of random classifications can occur does not preclude the presence of clustering. At high values of $d/N$, however, this method of pattern recognition would provide good classifications regardless of the presence of such clustering. In such a case, the data space should be examined by other methods in order to evaluate the actual class separations. Once separation of the classes is determined, discriminants can be used to simplify the problem. It should be noted that the ratios discussed above apply to classes containing equal numbers of patterns. The effects of unequal class sizes were discussed previously.[2]

## ACKNOWLEDGMENT

This research was supported by the U.S. Environmental Protection Agency under Cooperative Research Agreement CR 807531. The contents do not necessarily reflect the views of the Agency, and no official endorsement should be inferred. The PRIME 750 computer used in these studies was purchased, in part, with the support of the National Science Foundation.

## REFERENCES AND NOTES

(1) Stuper, A. J.; Jurs, P. C. *J. Chem. Inf. Comp. Sci.* **1976**, *16* (4), 238.
(2) Varmuza, K. "Pattern Recognition in Chemistry"; Springer-Verlag: New York, 1980.
(3) Tou, J. T.; Gonzalez, R. C. "Pattern Recognition Principles"; Addison-Wesley: Reading, MA, 1974.
(4) Stouch, Terry R.; Jurs, Peter C. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 45-50.
(5) Schrage, Linus *Assoc. Comput. Mach. Trans. Math. Software* **1979**, *5*, 132.
(6) Muller, Mervin E. *J. Assoc. Comput. Mach.* **1959**, *July*, 376.
(7) Moriguchi, Ikuo; Komatsu, Katsuichiro; Matsushita, Yasuo *J. Med. Chem.* **1980**, *23*, 20.
(8) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. "Computer Assisted Studies of Chemical Structure and Biological Function"; Wiley-Interscience: New York, 1979.
(9) Strang, Gilbert "Linear Algebra and Its Applications", 2nd ed.; Academic Press: New York, 1980.
(10) Belsley, David A.; Kuh, Edwin; Welsh, Roy E. "Regression Diagnostics: Identifying Influential Data and Sources of Collinearity"; Wiley: New York, 1980.
(11) Henry, Douglas R.; Jurs, Peter C.; Denny, William A. *J. Med. Chem.* **1982**, *25*, 899–908.
(12) Rose, Susan L.; Jurs, Peter C. *J. Med. Chem.* **1982**, *25*, 769–776.
(13) Stouch, Terry R.; Jurs, Peter C. *EHP, Environ. Health Perspect.*, in press.